Supplementary Figures - Calls

Figure S1. Example of a Somatic variant (S). Somatic calls are made when the variant has sufficient support in the tumor track with absence of obvious sequencing artifacts. In this example, the variant is presumed to be a real somatic variant. When evaluating the reference sequence in the Genome Features section, the reference allele is a cytosine (C). The alignments and coverage in the DNA tumor track show that approximately 20% of reads support a variant adenine (A) allele (green). Importantly, there are no reads supporting the variant in the normal sample, indicating that the variant is a somatic variant rather than a germline polymorphism. Using the gene annotation track, we can predict that this (C>A) base change would result in an ATG (M; Methionine) to ATT (I; Isoleucine) missense variant in the *PANK4* gene (Note: this gene is transcribed on the negative strand).

Helpful Hints:
1) Somatic variants, due to impure tumor samples, will typically have VAF less than 50%. However, the latter is not a strict rule because random sampling, copy-number alterations, loss of heterozygosity, and other factors can sometimes produce somatic VAF at or above 50%.
2) If the expected variant is not visualized during manual review, it is possible that: 1) IGV is not focused on the correct coordinates, 2) the genome version is incorrect, or 3) the supporting reads have been lost due to downsampling.

Figure S2. Example of a Germline variant (G). Germline calls are made when the variant has sufficient support in the normal track beyond what is considered attributable to tumor contamination of the normal. In this example, the variant is presumed to be a germline polymorphism. The reference allele is a guanine (G), however reads in the DNA normal/tumor tracks support a thymine (T) allele. This indicates that the variant is likely a homozygous germline polymorphism. The Single Nucleotide Polymorphisms (SNPs) Track provides further support that the variant in question is a common polymorphism.
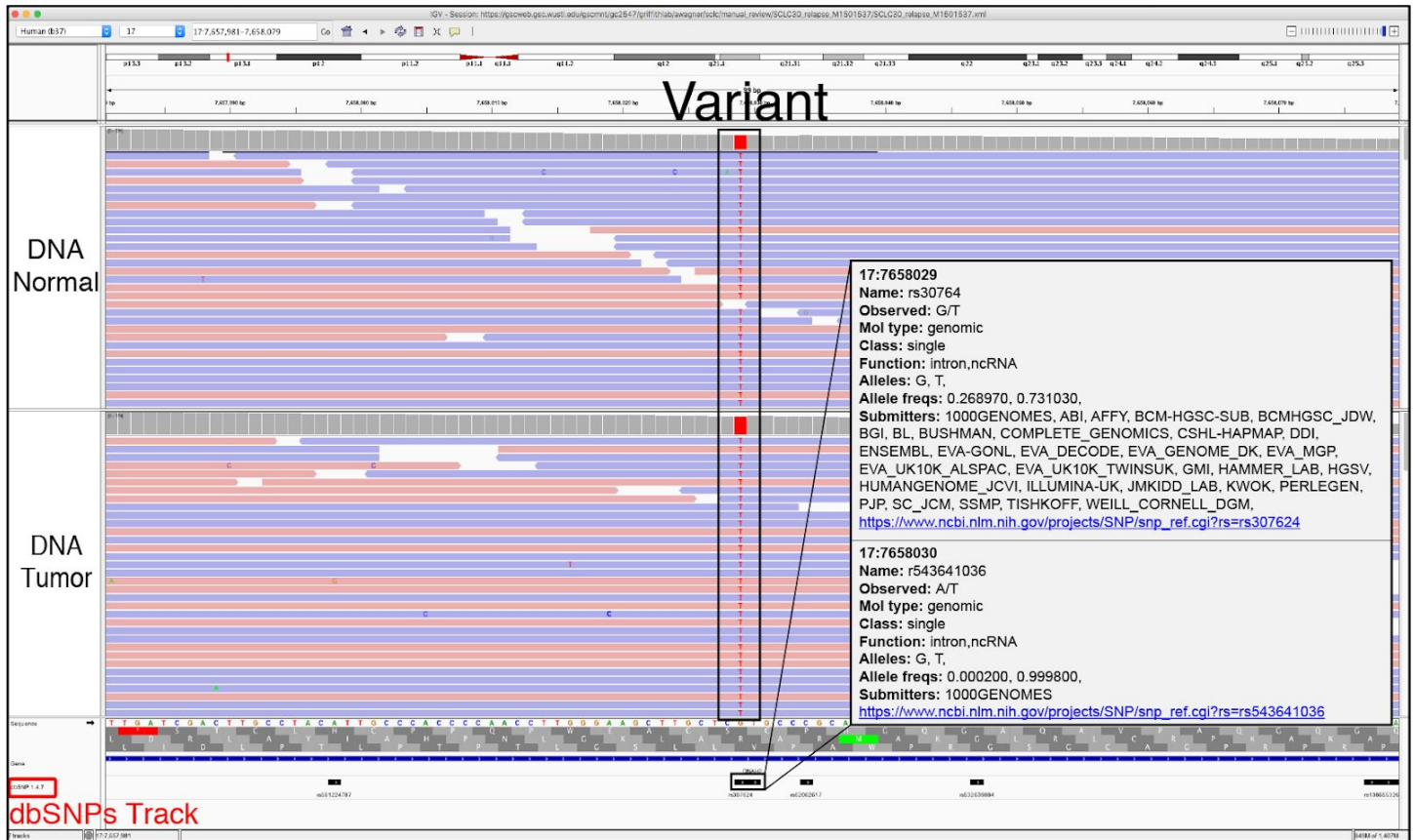
Helpful Hints:

1) Typically, germline variants present with a Variant Allele Frequency (VAF) near 50% or 100%, indicating hetero- or homozygosity, respectively.

2) Bulk tumors typically contain some normal cells. Therefore, given adequate depth, 100% VAF in a non-purified tumor sample should be suspicious and is likely a homozygous germline polymorphism.

3) To view the SNPs Track in the Genome Features section, use the "Load from Server" feature in IGV. Examples for loading this track are shown below:

    GRCH37: "File" > "Load from Server..." > "Annotations" > "Variation and Repeats" > "dbsnps 1.4.7"
    GRCH38: "File" > "Load from Server…" > "Annotations" > "Common Snps 1.4.2"

4) If the variant in question is also in the SNPs Track, then it is most likely germline. Clicking on, or hovering over, the grey bar in the SNPs Track will create a popup with additional information about the germline SNP.

5) A germline call after somatic variant caller filtering is suspect and might reveal underlying issues with the analysis pipeline being used.

Figure S3. Example of an Ambiguous variant (A). Ambiguous calls are made when the variant in question could be a true somatic variant, but the reviewer is not confident due to sequencing features, genomic context, and/or, corresponding reads. In this example, the variant has support from fourteen reads, but most are on negative read strands (93%). Additionally, several of the supporting reads have multiple mismatches indicating potentially low-quality reads. More information would be required to call this variant somatic or fail, therefore, the correct label is ambiguous.
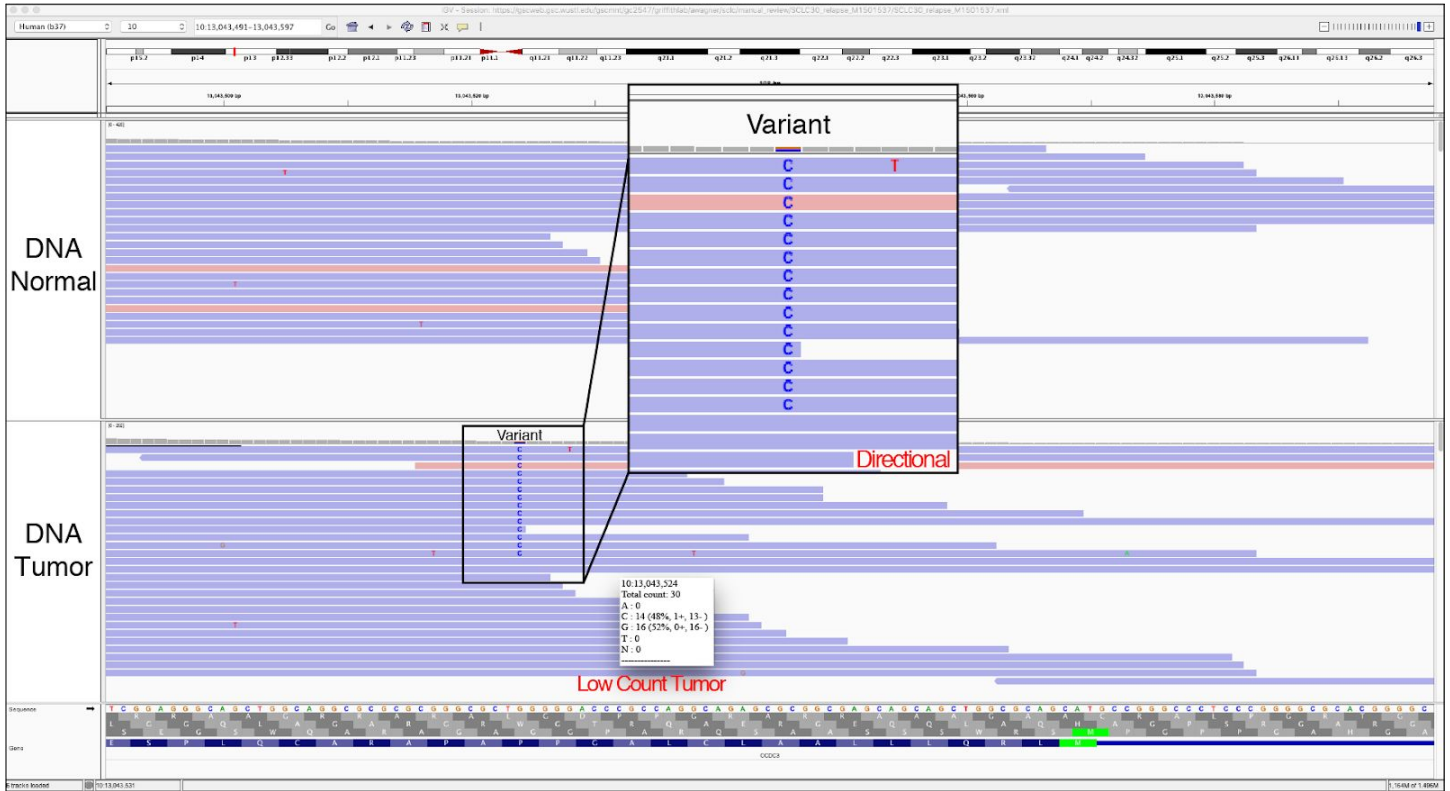
Helpful Hints:
1) Using Tags and Notes can help individuals understand why variants were labeled as ambiguous.

Figure S4. Example of a Failed variant (F). Failed calls are made when the variant has low variant support and/or reads that indicate a sequencing artifact. In this example, the variant in question is likely attributable to a pipeline artifact and is therefore not a true variant. When the IGV window is zoomed in, the variant appears to be somatic; however, in the provided zoomed-out window, we reveal a region of high discrepancy. High discrepancy regions (HDR) can suggest improper alignment in regions of high homology across the genome or errors in the reference assembly. Given the HDR pattern observed, this variant is most likely a false positive and should be failed during manual review.
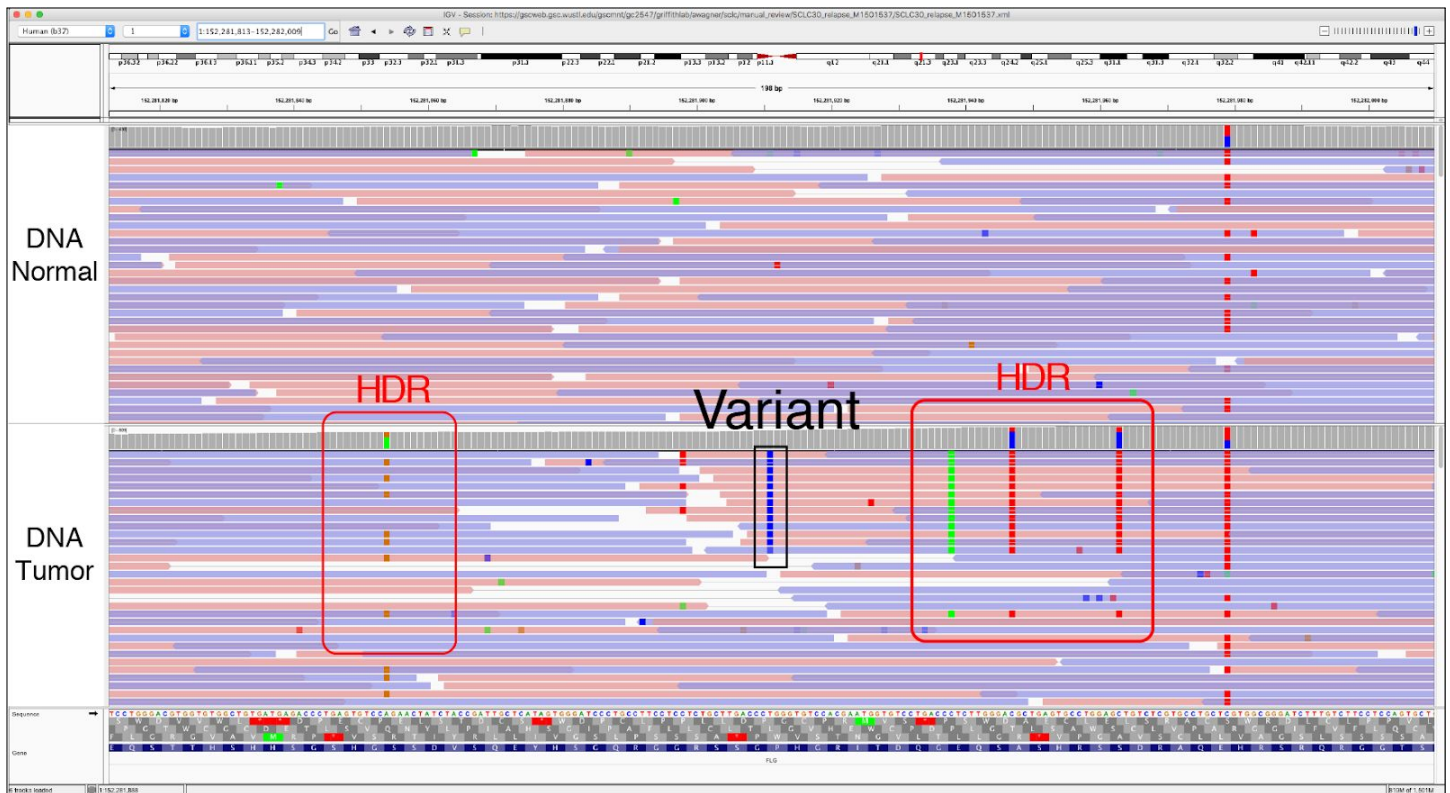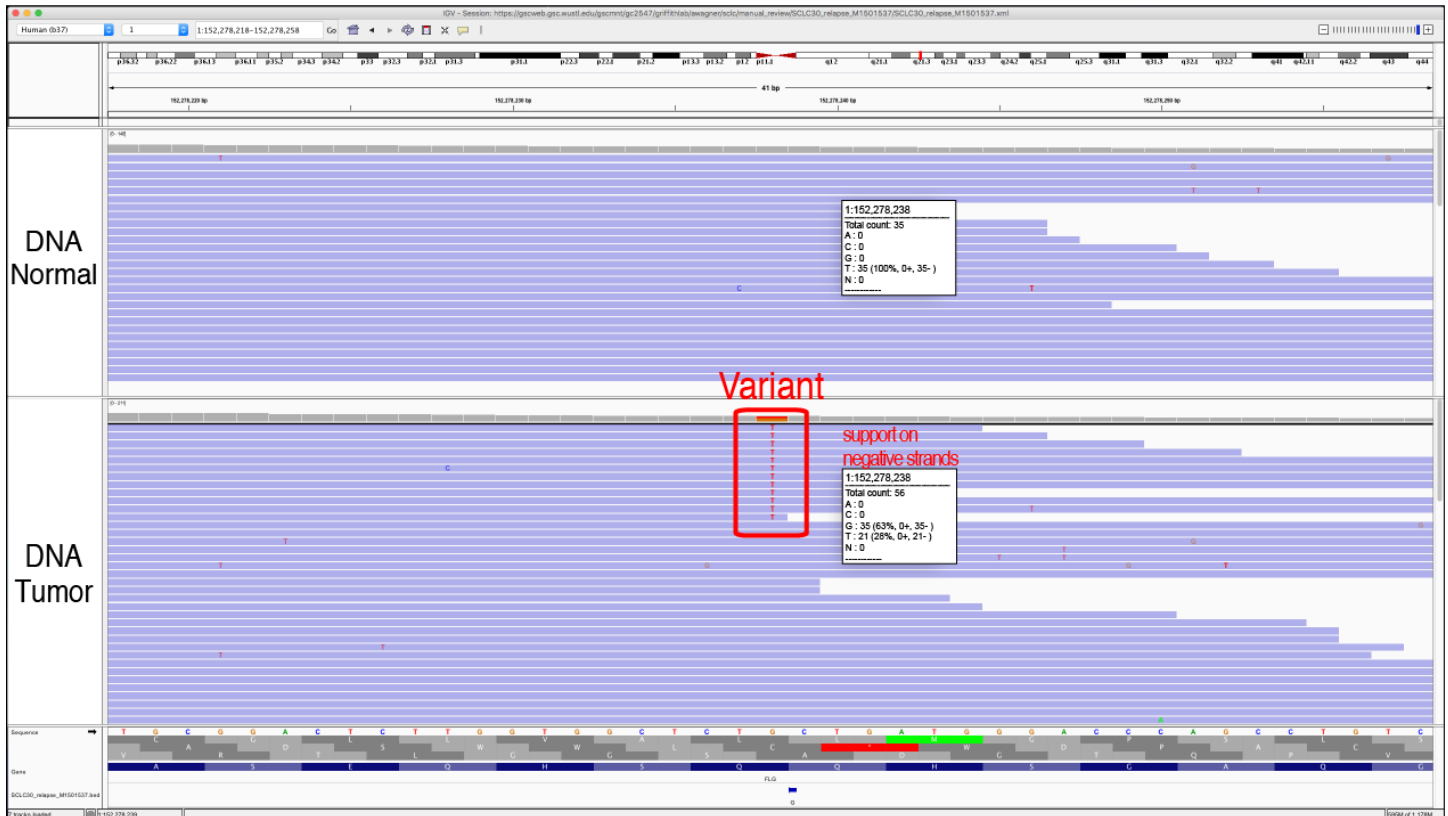

Helpful Hints:
1) Using Tags and Notes can help individuals understand why variants were labeled as fail.

Supplementary Figures - Tags

Figure S5. Example of Directional (D). The Directional tag is used when the variant in question can only be found on reads that are sequenced in either the positive or the negative direction. Typically, this is caused by strand bias during sequencing. To properly visualize the directional artifacts, IGV tracks must be colored by read strand.

Helpful Hints:

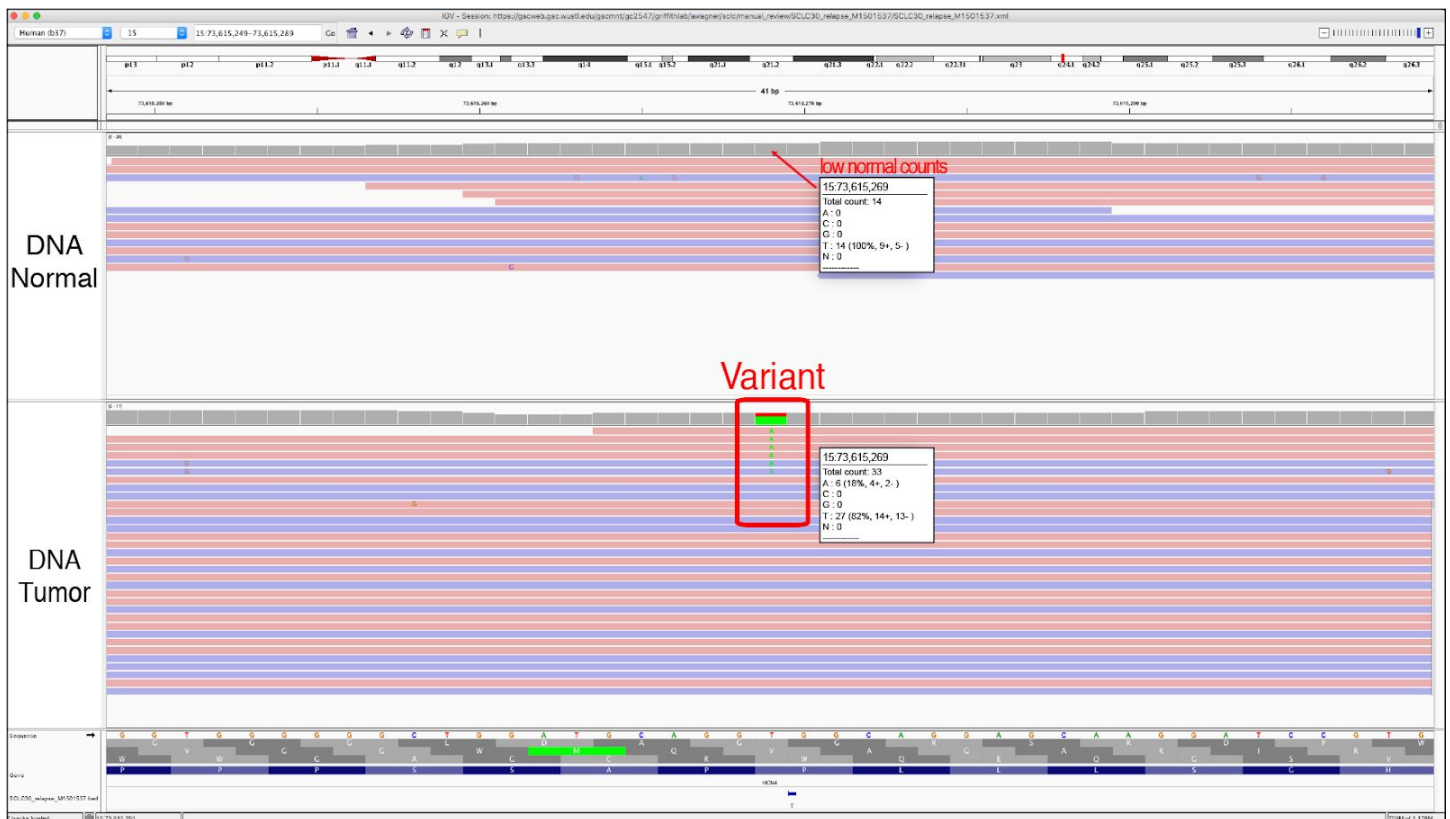1) This tag can best be assessed when the reads are not viewed as pairs. When viewing data tracks as pairs, the reads in both directions are overlaid and could possibly make the variant appear to be exclusively supported by read strands in a particular direction.

2) To observe this artifact, it is necessary to color the alignments by read strand:

> Right click on data track > "Color alignments by" > "read strand"

Figure S6. Example of No Count Normal (NCN). The No Count Normal tag is used when there is no coverage in the normal track, preventing adequate comparison to the tumor track. This can occur when there is no normal track available or if there is no coverage in the normal track at the locus in question. Typically, at least 20X coverage in both normal and tumor tracks is required to make accurate calls; however, this threshold is experiment-specific.

Helpful Hints:

1) If a variant has low coverage in the normal track, it can be treated like a tumor only sample. This might require populating the Genome Features section with a SNPs Track (e.g., dbSNP, 1000 genomes, ExAC, gnomAD, etc.) to ensure that the variant is not a polymorphism (see Step 3 in **Figure 3A** for setting up manual review).

2) Thresholds can be used to pre-filter variants with no coverage in tumor or normal to eliminate the need to evaluate these variants during manual review.

Figure S7. Example of Low Count Normal (LCN). The Low Count Normal tag is used when there is inadequate coverage in the normal track (coverage < 20X), preventing adequate comparison to the tumor track. A popup window with coverage information can be viewed by clicking on the locus position in the coverage track. Typically, at least 20X coverage in both normal and tumor tracks is required to make accurate calls; however, this threshold is experiment-specific.

Helpful Hints:

1) If a variant has low coverage in the normal track, it can be treated like a tumor only sample. This might require populating the Genome Features section with a SNPs Track (e.g., dbSNP, 1000 genomes, ExAC, gnomAD, etc.) to ensure that the variant is not a polymorphism (see Step 3 in **Figure 3A** for setting up manual review).

2) Thresholds can be used to pre-filter variants with low coverage in tumor or normal to eliminate the need to evaluate these variants during manual review.

Figure S8. Example of Low Count Tumor (LCT). The Low Count Tumor tag is used when there is inadequate coverage in the tumor track (coverage < 20X), preventing adequate comparison to the normal track. A popup window with coverage information can be viewed by clicking on the locus position in the coverage track. Typically, at least 20X coverage in both normal and tumor tracks is required to make accurate calls; however, this threshold is experiment-specific.
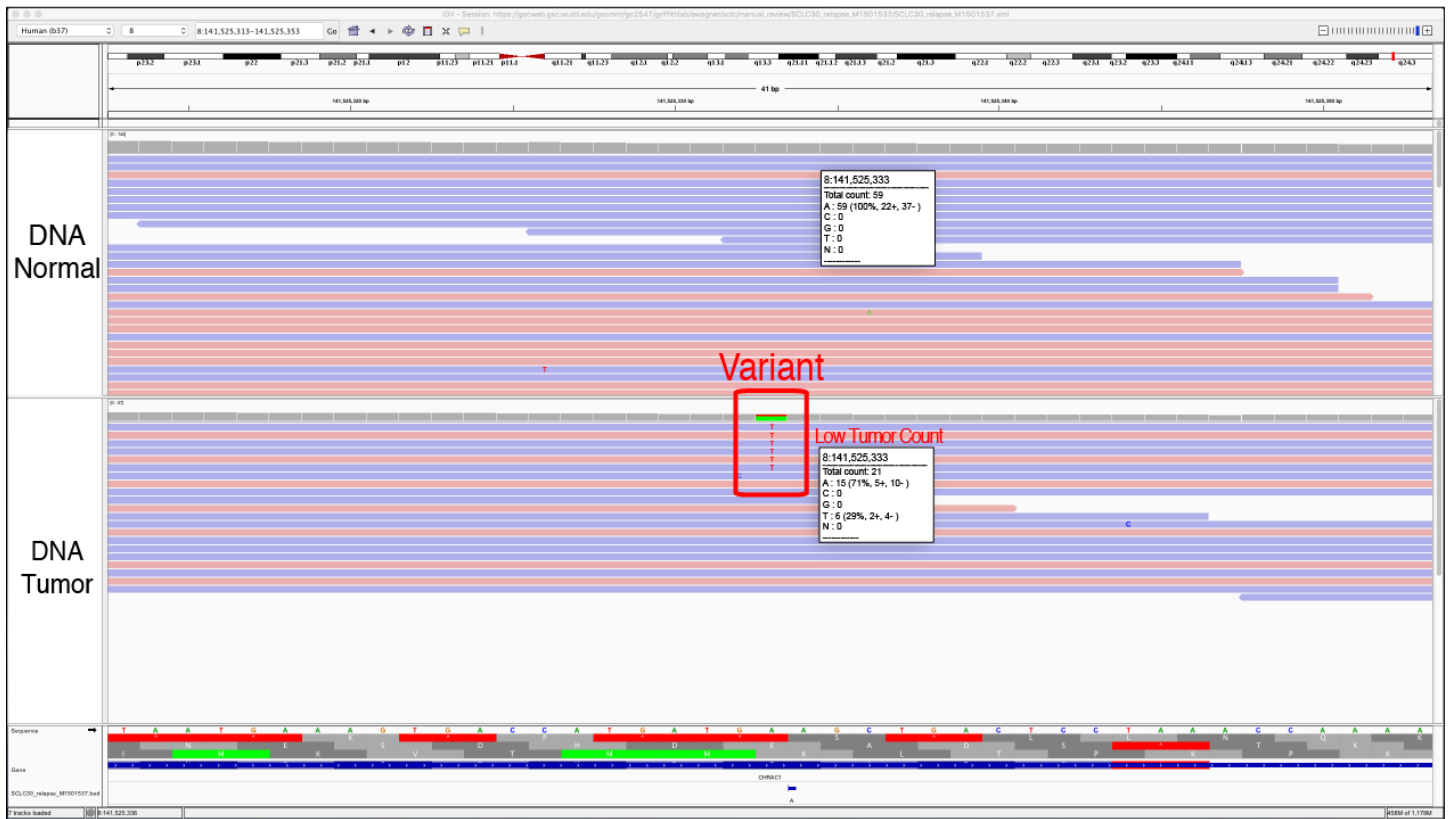
Helpful Hints:
1) Calling a variant with low coverage has important downstream implications. When the tumor track has low coverage, variant allele frequency (VAF) estimates can be heavily influence by sequencing noise and sampling bias. This may result in: a false negative with an underestimated VAF, a false positive due to over-estimation of the VAF, and/or a true positive call with inaccurate VAF.
2) The LCT tag acts as a bare minimum for tumor coverage but only in concert with a 5% VAF minimum with at least 4-5 reads of support (taking into account short inserts). Therefore, the LCT tag can denote that a variant was considered ambiguous or somatic in a rare sequencing context.
3) Thresholds can be used to pre-filter variants with low coverage in tumor or normal to eliminate the need to evaluate these variants during manual review.

Figure S9. Example of Multiple Variants (MV). The Multiple Variants tag is used if the variant's locus has reads supporting three or more different alleles. In the example shown, there is read support for all four nucleotides (A, C, G, and T) at the same locus. If the putative variant base co-occurs with multiple instances of other bases, it is less likely to be a true somatic variant.
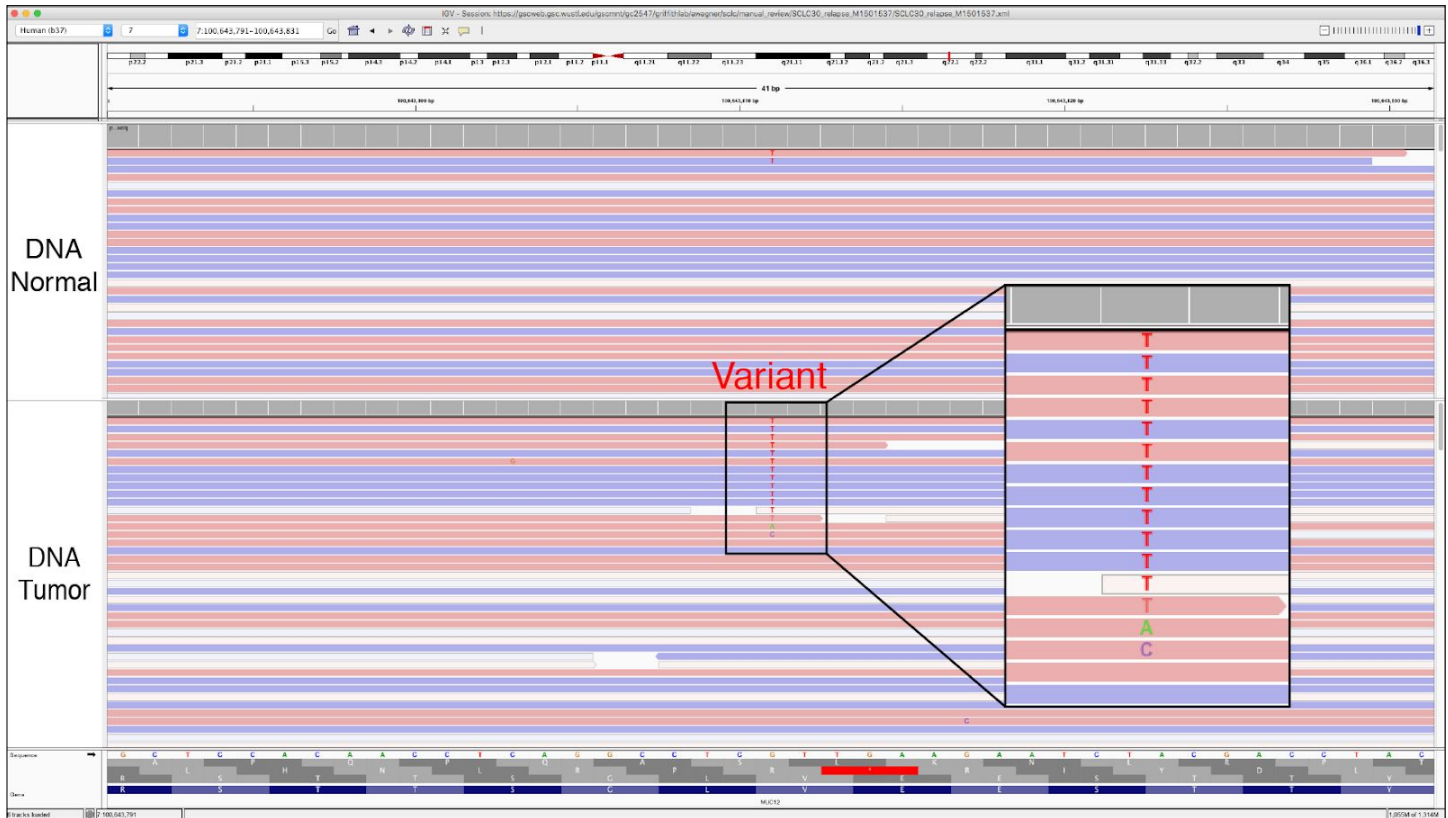
Helpful Hints:

1) Clicking on the coverage track will reveal a popup window with relative abundance of each base at the selected locus.

2) Do not rely on coverage track coloring as there might be multiple variants that have a variant allele frequency (VAF) too small to be represented in the coverage bar. The VAF threshold for coloring the coverage bar can be changed in the IGV preferences panel:

"View" > "Preferences" > "Alignments" > "Coverage allele-fraction threshold" > insert threshold

3) For very deep data, multiple variants due to random error will start to accumulate. The relative abundance of each base should be considered in cases with deep coverage.

4) While rare, true multi-allelic somatic variants are possible in tumors.

Figure S10. Example of Low Variant Frequency (LVF). The Low Variant Frequency tag is used when there are some reads of support for the variant, but the variant allele frequency (VAF) is relatively low. A popup window with VAF information can be viewed by clicking on the locus position in the coverage track. Typically, at least 5% VAF is required to make confident calls (given 20X coverage); however, this threshold is experiment-specific.

Helpful Hints:
1) The coverage track will be colored according to base when a variant is present at the default VAF. This threshold can be changed in the IGV preference panel:

    "View" > "Preferences" > "Alignments" > "Coverage allele-fraction threshold" > insert threshold

    This can be particularly helpful for high depth samples and/or when low VAF (e.g., sub-clonal) variants are expected. With sufficient depth of coverage, the VAF threshold can be reduced.
2) Thresholds can be used to pre-filter variants with low tumor VAF to eliminate the need to evaluate these variants during manual review.

**Figure S11. Example of Multiple Mismatches (MM).** The Multiple Mismatches tag is used when the reads that contain the variant have other mismatched base pairs, which reduces the confidence in the read quality. Specifically, given a high error rate and a random distribution of errors, spurious variants can occur when the errors align across reads in the tumor sample but not in the normal sample. The MM and HDR tags are similar, in that both relate to mismatches in reads containing the variant; however, the MM tag is used when multiple mismatches are distributed unevenly (see **Figure S12**).
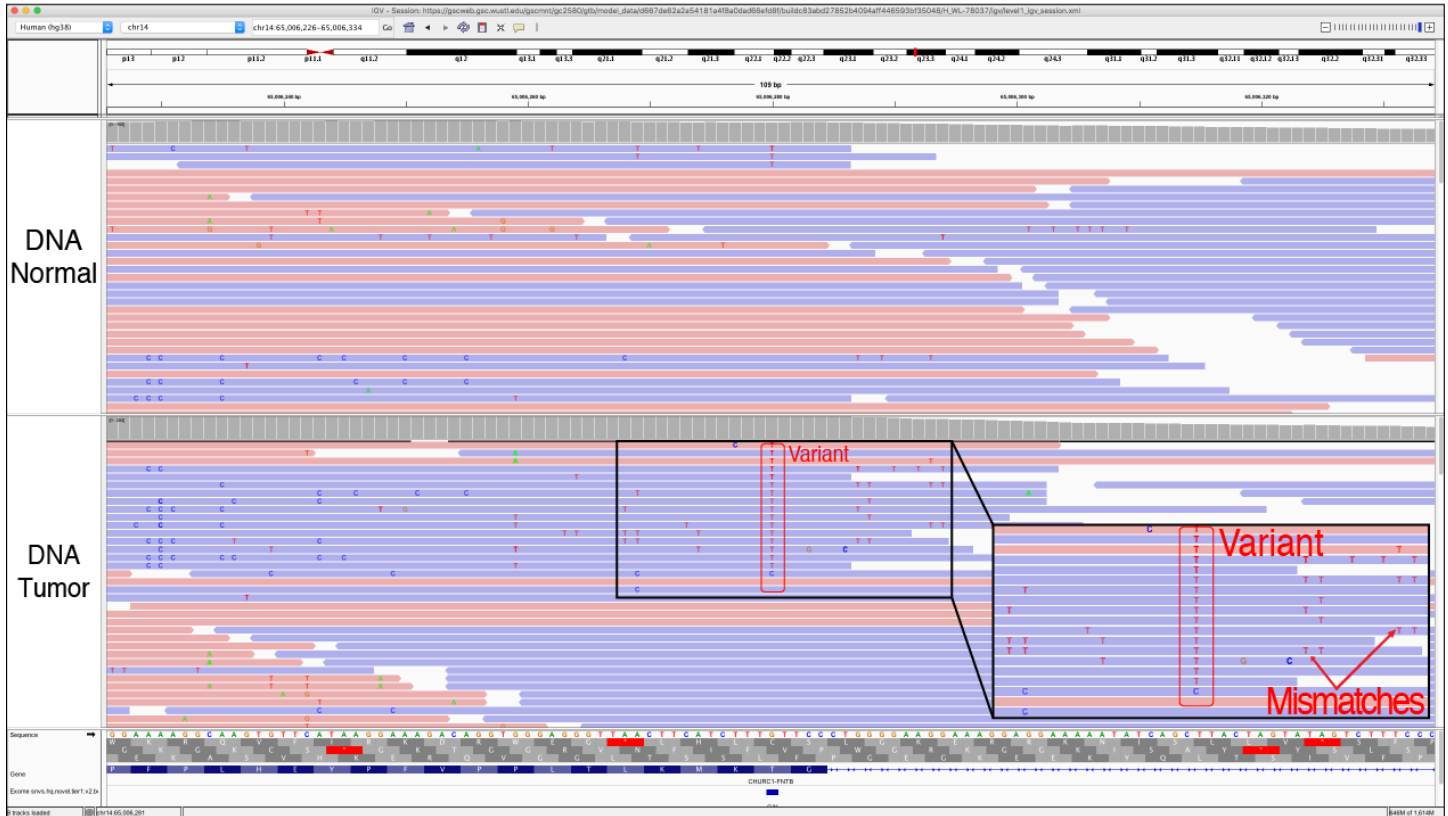
Helpful Hints:

1) If the mismatches are of high quality, this likely indicates that the read was properly sequenced. In this case, the mismatches occur due to misalignment. If the mismatches are of low quality, this likely indicates that the read was improperly sequenced. Both of these examples reduce confidence in the variant.

2) High densities of mismatches in the tumor track increase the probability that identical base substitution errors align across reads causing the VAF to surpass filtering thresholds. The higher the read depth, the less likely this situation is to arise, as low percentage VAF variants increase in plausibility with increased read depth.

Figure S12. Example of High Discrepancy Region (HDR). The High Discrepancy Region tag is used when most reads containing the variant also contain other mismatches at the same locus. Typically, HDRs are observed when reads map to incorrect but homologous regions that contain localized differences, which are interpreted as variants. The HLA loci, duplicated loci, and other highly polymorphic regions are especially prone to this issue. These regions may require specialized alignment or assembly strategies for high quality variant calling.
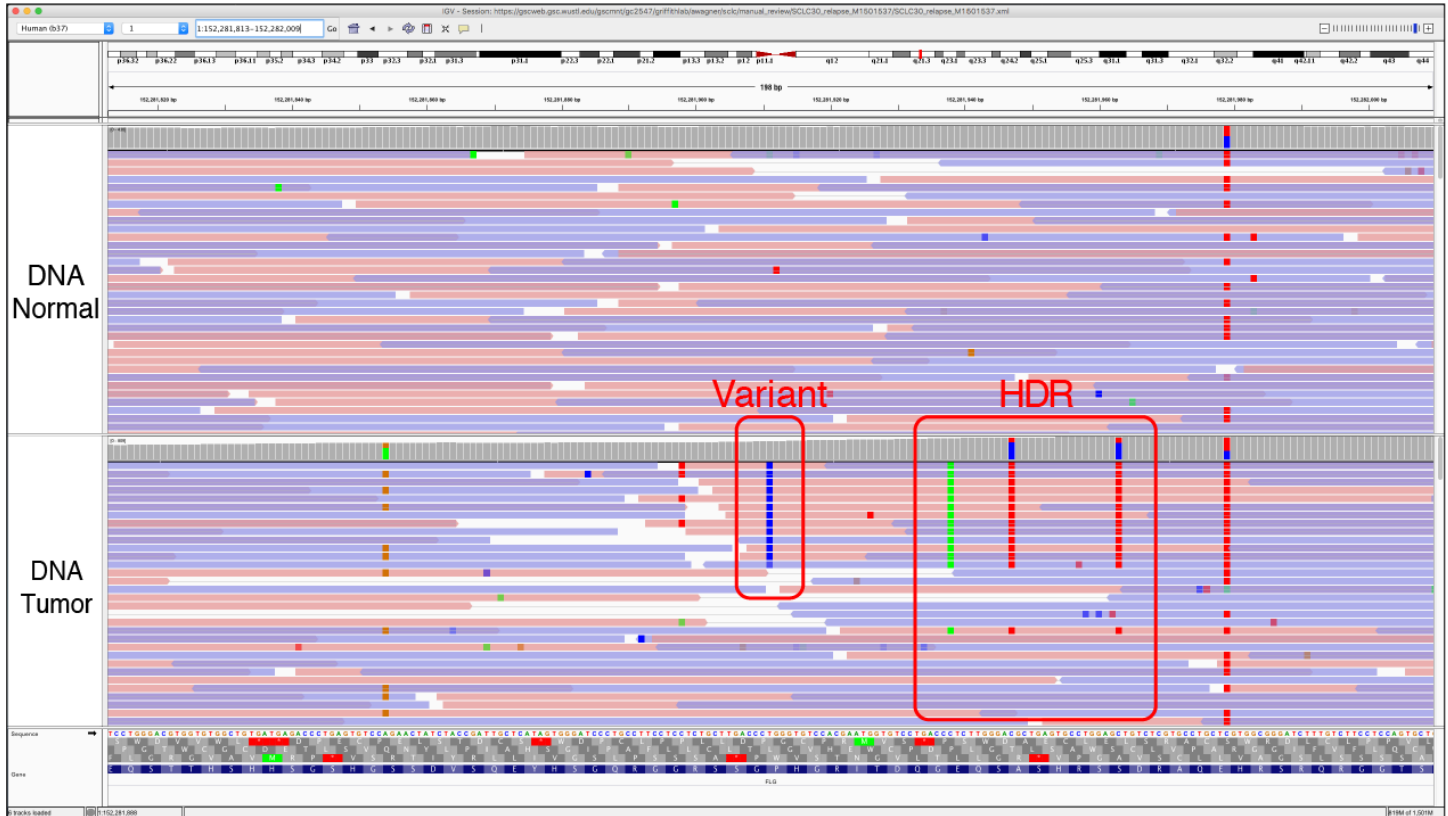
Helpful Hints:

1) The presence of more than three identical mismatches within a 100-200 base-pair region is highly indicative of an HDR.

2) It is important to be sure that the variant being evaluated is not surrounded by a cluster of single nucleotide polymorphisms (SNPs). Sometimes, true variants can occur in close proximity to multiple SNPs and might be confused with an area of HDR. This is particularly true for individuals with haplotypes that are not well-represented by the reference sequence.

Figure S13. Example of Low Mapping (LM) quality. The Low Mapping tag is used to indicate variants that are mostly supported by reads that have low mapping quality. When reads are colored by readstrand, translucent/transparent reads indicate lower mapping quality and opaque reads indicate higher mapping quality. Mapping quality refers to a measure of confidence or probability that a read has been correctly aligned to the reference genome. Variants that are supported primarily or solely by low mapping quality reads are considered suspect.
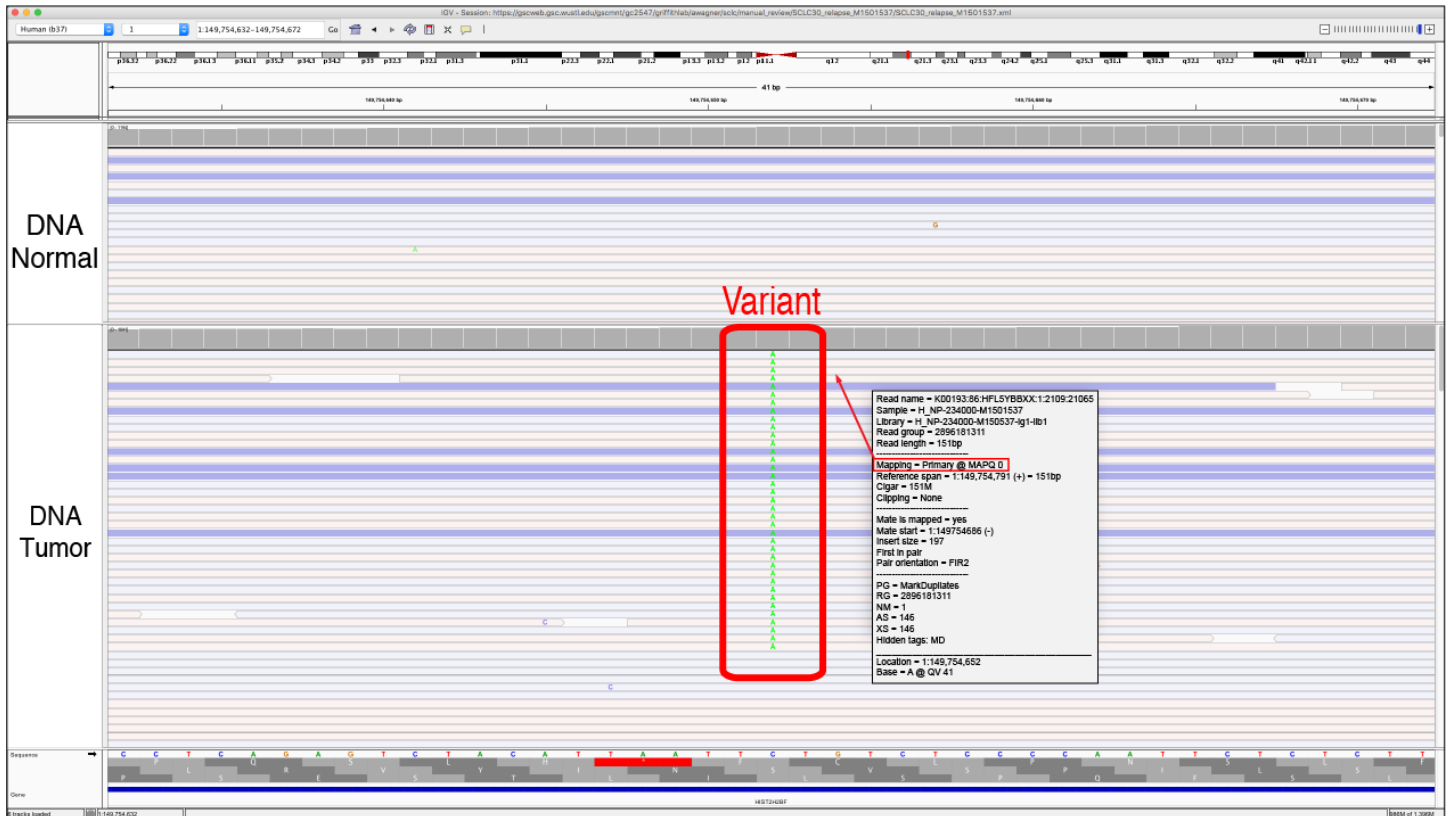
Helpful Hints:

1) Mapping quality scores can be ascertained by clicking on the read.
2) In regions where numerous reads have a mapping quality of 0, the reads are often mapped to multiple locations across the genome. This results in low mapping quality reads in both the normal and tumor tracks. Alternate mapping locations can be ascertained by clicking on the read.
3) By default, all reads are shown in IGV, even if the mapping quality is 0. This threshold can be adjusted to eliminate low quality reads from IGV during manual review:

    "View" > "Preferences" > "Alignments" > "Mapping quality threshold" > insert threshold

Figure S14. Example of Tumor in Normal (TN). The Tumor in Normal tag is used to indicate that the variant has reads of support in the normal track. This is a common occurrence in certain blood tumors (e.g., leukemia) as well as tumors that are highly metastatic. In some instances, TN might be a reason to fail the variant, whereas in other situation it can be used to denote ambiguity in the manual review call.
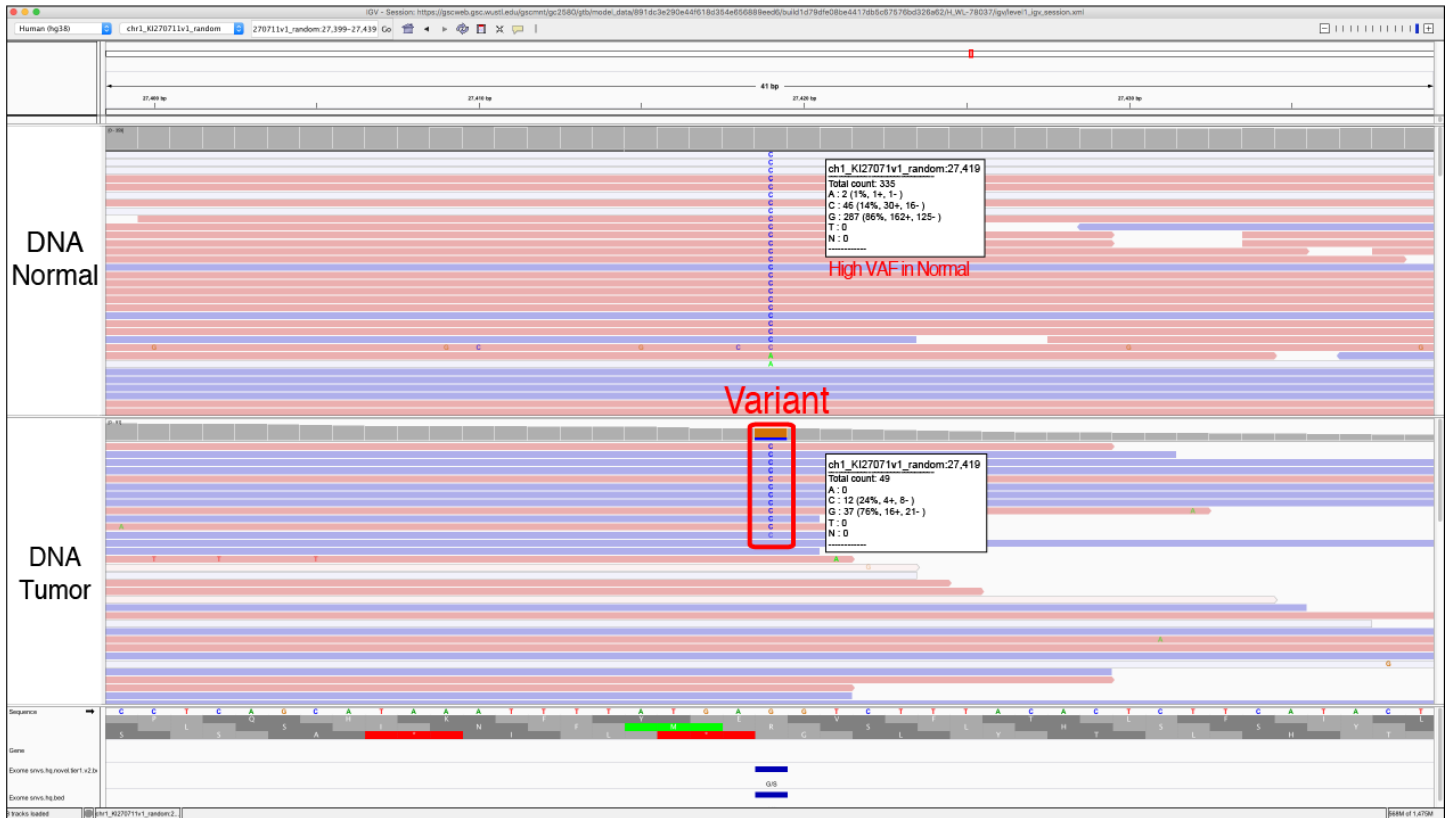
Helpful Hints:

1) TN does not occur in all hematopoietic tumors but is likely when tumor cells are circulating in the bloodstream (e.g., acute myeloid leukemias with high blast counts).

2) Tumors that are metastatic may have tumor cells circulating in the bloodstream and thus can also have TN contamination.

3) Problems with sample barcoding (indexing) or cross contamination of samples can also lead to apparent support for a somatic variant in the normal.

4) Evaluating other normal samples from your cohort, or evaluating multiple variants within the same sample/experiment, can help set a relative acceptable TN threshold. This will help to differentiate sequencing and pipeline artifacts from tumor contamination of normal tracks.

5) Variants created by sequencing or alignment artifacts will also often occur in both the tumor and the normal tracks and can be labeled with TN.

Figure S15. Example of Short Inserts (SI) and Short Inserts Only (SIO). The Short Inserts tag is used when the variant is found on small nucleic acid fragments whereby sequencing from each end results in overlapping reads. In IGV, this is indicated as a grey bar through the middle of reads when reads are viewed as pairs. Variants supported by read pairs produced from these short fragments can result in the appearance of two independent reads supporting a variant when in reality, they represent only a single nucleic acid molecule. The SI tag is used when support for the called variant is primarily from short-insert read pairs but other read strands that are not short inserts also show variant support. The SIO tag is used when support for the called variant is exclusively present in paired reads from short inserts. This issue is prevalent in data derived from archival material (FFPE samples) or other source material with small/degraded DNA fragments (e.g., cell-free DNA).

Helpful Hints:

1) To visualize short insert variants you must view the tracks as pairs. Regions where the paired reads overlap will be dark purple and contain a horizontal grey line. At the ends, where there is no overlap, reads will remain blue or pink. Reads can be viewed as pairs using IGV commands:

    right click each data track > "View as pairs"

2) When viewing reads as pairs, short inserts can be observed; however, it will also overlay reads to reduce the total information available to the reviewer.

3) Short inserts are generally observed at lower variant frequencies and present in two or three read pairs (i.e., four to six reads in total).

Figure S16. Example of Adjacent Indel (AI). The Adjacent Indel tag is used when a somatic variant was possibly caused by misalignment around a germline or somatic insertion/deletion (indel). In this example, it is likely that a real somatic variant is present, however, the variant is neither a simple 'A' insertion, nor a simple 'A' substitution. It is possible that the true variant is an 'AA' insertion that was miscalled by the automated somatic variant callers.

Helpful Hints:

1) To effectively visualize this pattern, it is necessary to zoom out using the IGV Genome Ruler.
2) It is important to evaluate the Genome Features section to visualize possible tandem repeats that might be implicated in the misalignment.
3) These cases can sometimes be resolved by correcting the nature of one or more called variants rather than failing the variant entirely. This is an instance where the IGV Notes section would be valuable.
4) This phenomenon is common with larger deletions where ends of reads will be misaligned within the deletion.

Figure S17. Example of Same Start/End (SSE). The Same Start/End tag is used when the variant is only contained by reads that start and stop at the same genomic loci. This is typically attributed to a variant called in multiple reads created from the same originating molecule during the library amplification process but erroneously not removed during read deduplication.

Helpful Hints:
   1) Identifying SSE artifacts requires first sorting the reads by base and subsequently zooming out to view a larger genomic region. This allows for visualization of the ends of the reads.

Figure S18. Example of End of reads (E). The End of reads tag is used when the variant called is within 30 base pairs of the end of the variant-supporting reads. At read ends (especially the 3' end), there is an increased rate of error generation that can cause appearance of an erroneous variant.

Helpful Hints:

1) Identifying End of reads artifacts requires first sorting the reads by base and subsequently zooming out to view a larger genomic region. This allows for visualization of the ends of the reads.
2) Additional mismatches downstream the called variant can increase confidence that the variant in question is a sequencing artifact.
3) This artifact is more easily evaluated by coloring the alignments by read strand:

        Right click on data track > "Color alignments by" > "read strand"

Figure S19. Example of Mononucleotide repeat (MN). The Mononucleotide tag is used when a variant is called in proximity to a region of the reference sequence that contains a single nucleotide repeat (e.g., AAAAAAA…). In this instance, the called variant is most likely caused by misalignment of the reads to the reference genome. Some sequencers, particularly those dependent on the polymerase, are prone to making mistakes in repeat regions. However, it is important to note that mononucleotide repeats are also a common source of real human variation (inherited germline, de novo germline, or somatic) that arise due to errors produced by polymerase during DNA replication. Other factors, such as the size of the repeat, the VAF, or appearance in the normal, should be considered during manual review to confidently call the variant. The frequency in other samples processed in the same way (capture reagent, alignment algorithm, etc.) can be helpful in identifying common artifacts. Special alignment, assembly, or even additional sequencing technologies may be needed to validate short repeats of this nature.

Helpful Hints:

1) Typically, these variants are small deletions or insertions, and they are usually visualized in both the tumor and normal tracks.
2) Although the variant being evaluated here is a one base-pair deletion, other reads at the same locus typically have insertions and deletions of varying lengths.

Figure S20. Example of Dinucleotide repeat (DN). The Dinucleotide repeat tag is used when a variant is called in proximity to a region of the reference sequence that contains two alternating nucleotides (e.g., TGTGTG…). In this instance, the called variant is most likely caused by misalignment of the reads to the reference genome. Some sequencers, particularly those dependent on the polymerase, are prone to making mistakes in repeat regions. However, it is important to note that dinucleotide repeats are also a common source of normal human variation (inherited germline, de novo germline, or somatic) that arise due to errors produced by polymerase during DNA replication. Other factors, such as the size of the repeat, the VAF, or appearance in the normal, should be considered during manual review to confidently call the variant. The frequency in other samples processed in the same way (capture reagent, alignment algorithm, etc.) can be helpful in identifying common artifacts. Special alignment, assembly, or even additional sequencing technologies may be needed to validate short repeats of this nature.

Helpful Hints:
1) Typically, these variants are small deletions or insertions and they are usually visualized in the both tumor and normal tracks.
2) Although the variant being evaluated is a two base-pair deletion, other reads at the same locus typically have insertions and deletions in multiples of two.

Figure S21. Example of Tandem Repeat (TR). The Tandem Repeat tag is used when a variant is called in proximity to a region of the reference sequence that contains some number of repeated nucleotides (e.g., GTGGTGGTG…). In this instance, the called variant is most likely caused by misalignment of the reads to the reference genome. Some sequencers, particularly those dependent on a polymerase, are prone to making mistakes in repeat regions. However, it is important to note that tandem repeats are also a common source of normal human variation (inherited germline, de novo germline, or somatic) that arise because of errors produced by polymerase during DNA replication. Other factors, such as the size of the repeat, the VAF, or appearance in the normal, should be considered during manual review to confidently call the variant. The frequency in other samples processed in the same way (capture reagent, alignment algorithm, etc.) can be helpful in identifying common artifacts. Special alignment, assembly, or even additional sequencing technologies may be needed to validate short repeats of this nature.

Helpful Hints:

1) Typically, these variants are small deletions or small insertions and they are usually visualized in both the tumor tracks and the normal tracks.
2) In this example, the variant being evaluated is a three base-pair deletion, whereas other reads at the same locus have insertions and deletions in multiples of three, which reduces confidence in the called variant. This pattern can help distinguish a TR artifact from a true somatic variant.

Figure S22. Example of Ambiguous Other (AO). The Ambiguous Other tag is used to define a variant surrounded by inconclusive genomic features that cannot be explained by the other tags. In this example, we observe a low complexity region (e.g., genomic regions with increased A/T or G/C content), which can accurately be described with the AO tag.
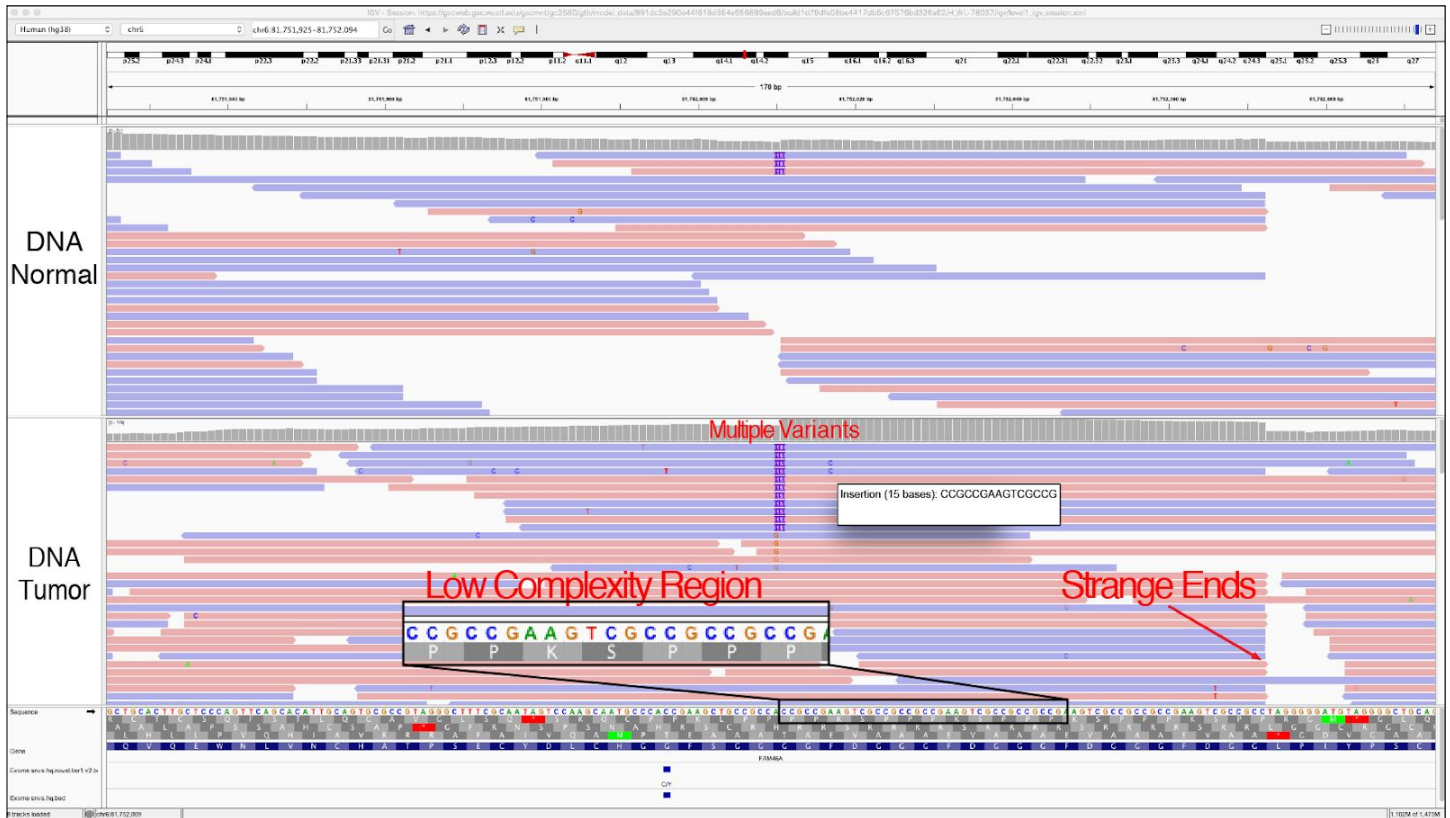
Helpful Hints:

1) If the Ambiguous Other tag is used, it is highly recommended to include a short description in the notes section.